White Paper Report

Report ID: 100476

Application Number: HG-50020-10

Project Director: Gregory Crane (gregory.crane@tufts.edu)

Institution: Tufts University

Reporting Period: 7/1/2010-6/30/2014

Report Due: 9/30/2014

Date Submitted: 10/21/2014

The Hellespont Project: Integrating the material and linguistic record of the past

An NEH White Paper
Gregory R. Crane
Tufts University
HG-50020-10
7/1/2010-6/30/2014

a. Project Activities

The Hellespont project at the Perseus Digital Library[1] (PDL) focused on a number of major activities during the course of the grant, namely: 1) improvement of text markup: the conversion of Text Encoding Initiative (TEI)[2] P4 texts to TEI P5 Unicode-compliant versions; 2) standardization of legacy data: the mapping of data from the Perseus Art & Archaeology (A&A) database[3] to the CIDOC-CRM data model;[4] 3) creation of standardized descriptions for data: the production of MODS/MADS "FRBR" data for classical authors and editions and the release of the Perseus Catalog;[5] 4) assistance in the creation of extensive documentation[6] regarding difficulties encountered in these tasks; 5) the refining and publication of the CIDOC-CRM data and the release of other Perseus data as either RDF[7] linked data[8] or as "linkable" data with permanent URIs. Each of these activities will be discussed briefly in turn below.
The conversion of texts within the PDL from P4 to P5 Unicode compliant versions is still an ongoing task although much work was conducted during the course of this project. In the first year of the grant, some time was spent exploring the possibilities of converting the Perseus P4 texts to the TEI Analytics format[9] developed by the MONK Project but this was eventually abandoned in favor of utilizing the growing Epidoc standard[10] instead.
One of the most time consuming activities of the grant proved to be mapping the Perseus A&A data to the CIDOC-CRM as it involved several staff members and frequent communication between the two project teams. While there had originally been some consideration of mapping the Perseus and Arachne schemas to each other, as the Arachne project had recently completed a mapping of their data to the CIDOC-CRM model, it was decided that mapping Perseus A&A data to CIDOC and then aggregating the CIDOC-CRM data of both collections would be more productive. The Perseus A&A database contains nearly 20 years of legacy data

---

[1] http://www.perseus.tufts.edu/hopper/
[2] http://www.tei-c.org/index.xml
[3] http://www.perseus.tufts.edu/hopper/artifactBrowser
[4] http://www.cidoc-crm.org/ The **CIDOC Conceptual Reference Model (CRM)** provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. CIDOC is the International Committee for Documentation for the International Council of Museums.
[5] http://catalog.perseus.org
[6] http://hellespont.dainst.org/startpage/docu.html
[7] http://www.w3.org/RDF/
[8] http://en.wikipedia.org/wiki/Linked_data
[9] http://wiki.tei-c.org/index.php/TEI-Analytics
[10] http://sourceforge.net/p/epidoc/wiki/Home/

covering nearly 6,000 artifacts and staff review found many inconsistencies in the data and documentation. In many instances, there were no easily equivalent classes found within the CIDOC CRM to represent Perseus data. For example, there are many different ways to express key notions of time and place in CIDOC CRM, and it was discovered that the A&A database had a number of different ways of expressing these concepts as well. The Perseus A&A database also used a number of fields inconsistently such as those for data types, bibliography and locations. A number of tools were used to allow staff to collaboratively work on the mapping of the data, including a Google spreadsheet to track mapping of individual A&A data fields and list potential CIDOC-CRM equivalents and notes, the Arachne database mapping to the CIDOC-CRM (available both as one large XML file and as an online browsing tool), the CIDOC-CRM documentation itself, and a number of examples where Arachne staff created mappings for items shared by one or both databases. Ultimately all of the data fields in A&A were successfully mapped to the CIDOC CRM and the mappings were reviewed by our German colleagues. The final steps involved Perseus project staff developing an XSLT stylesheet from the mappings in the spreadsheet to transform the data into CIDOC-CRM.

The creation of "FRBR-inspired"[11] MODS[12] and MADS[13] records occurred throughout the entirety of the grant period. Although the original idea was to make these records available through the ZENON bibliographic database[14] of the German Archaeological Institute (DAI) it was ultimately determined that a separate catalog instance that could take full advantage of the fact that the catalog and authority records were created in XML. A prototype catalog implementation originally made use of the eXtensible catalog[15] open source software project, but the project decided ultimately to use Blacklight[16], an open source project that utilizes Ruby on Rails[17] and provides discovery interfaces for Solr[18] indexes. In addition, the catalog also makes use of CTS[19] (Canonical Text Services) URNS that compile into CITE (Collections, Indices, Texts and Extensions) tables[20]. All of this work has been conducted with the ultimate goal of releasing all of the data within the catalog as full linked data[21] while also making all of the MODS and MADS records available through GitHub.[22] One unexpected additional outgrowth of this work has been the inclusion of the Perseus created MADS records as part of the Virtual International Authority File (VIAF) in an effort to expand name variants, particularly those in Ancient Greek.[23]

---

[11] FRBR = Functional Requirements for Bibliographic Records, http://www.loc.gov/cds/downloads/FRBR.PDF

[12] Metadata Object Description Schema, an XML-based bibliographic description schema.

[13] Metadata Authority Description Schema, an XML schema that provides an authority element set to complement MODS.

[14] http://www.dainst.org/en/zenon

[15] http://www.extensiblecatalog.org/

[16] http://projectblacklight.org/

[17] http://rubyonrails.org/

[18] http://lucene.apache.org/solr/

[19] http://www.homermultitext.org/hmt-doc/cite/texts/ctsoverview.html

[20] http://sites.tufts.edu/perseuscatalog/documentation/user-guide/catalogdata-uris/

[21] For more on these future plans, please see: http://sites.tufts.edu/perseuscatalog/?p=170

[22] https://github.com/PerseusDL/catalog_data

[23] For more on this project, please see http://www.oclc.org/research/activities/viaf-scholars.html.

As much of the work conducted during this grant was collaborative, iterative, and highly technical in nature, detailed documentation was created to highlight key points and implementation details. The main body of documentation is available at the project website[24] and covers entity annotation, the creation of linked data, details on the CIDOC CRM mappings and aggregation, and details on the creation of various interfaces to the data. Documentation for the Perseus Catalog is available at the catalog project website[25] as well as a user guide.

Much of the final activity for this grant involved the refinement and several review iterations of the Perseus A&A CIDOC Mappings and making all of this data available as RDF.[26] In order to output this data as RDF/XML this required stable URIs for references and vocabulary terms within the Perseus data, although the nature of that data made it impractical to make all of the subjects and objects within the Perseus RDF have addressable URIs. In practice, the project mapped to existing URIs (such as the use of Pleiades[27] URIs for place names) whenever possible but did have to construct some new URI identifiers for some controlled vocabulary terms in this data, in particular date periods and artifact types. Other final work in the creation of Perseus linked data and cross-linking between the various Hellespont resources involved making all types of data found within the PDL itself linkable. This required additional planning regarding how to construct URIs for various resources including primary source texts and catalog data so that they are stable and dereferencable. The project began to publish stable URIs for various pre-existing resources in the library[28] and also used CTS Passage URNs in combination with URI prefixes to identify text passage to support linking both to texts as a whole and to individual passages. This involved the modification of the markup structure of many Perseus texts in order to support the retrieval of XML for individual text passages, all of which has informed the ongoing work converting texts to TEI P5 and Unicode.

In this last year of the Hellespont project, following a grant extension, work for the grant involved further work on the Perseus catalog by project staff as well as staff travel and participation and or presentation at various conferences and workshops to discuss and promote work conducted during this grant. This travel included staff attendance at the HathiTrust Unconference in September of 2013 in Illinois and partly to attend the Classical Association Conference in the United Kingdom in April of 2014. Project staff also attended and presented on the Perseus Catalog at a May workshop at the Center for Hellenic Studies (CHS) entitled "Libraries and the Open Greek and Latin Project." Grant support also enabled a separate day of discussions with Leipzig team members in advance of the NEH-funded IATDH workshop.[29]

In addition, the project PI presented the results and publications related to this project at the TEI annual conference in October 2013 in Rome, the AIA/APA conference in January of 2014, at

---

[24] http://hellespont.dainst.org/startpage/docu.html

[25] http://sites.tufts.edu/perseuscatalog/documentation/

[26] RDF downloads are available for each individual artifact category form the A&A database (namely buildings, coins, gems, sculptures, sites and vases) from the main webpage (http://www.perseus.tufts.edu/hopper/artifactBrowser). The aggregated RDF data can be downloaded here: http://hellespont.dainst.org/startpage/files/hellespont-partners-crm-rdf.tgz

[27] http://pleiades.stoa.org/

[28] For more detail on this process please see http://sites.tufts.edu/perseusupdates/beta-features/perseus-stable-uris/

[29] http://sites.tufts.edu/digitalagetext/2014-workshop/schedule/

two workshops at the CHS in May and June of 2014, and the DATECH conference in Madrid in May of 2014. In particular, the project PI had organized the May CHS meeting to discuss the Perseus catalog, the creation of stable identifiers and URIs for classical authors and texts, and potential ways that the libraries represented at the workshop could contribute metadata to the catalog and scans for classical editions for the growing Open Greek and Latin Project at Leipzig. Publicity efforts for this grant and its results largely involved the creation of two project websites at the DAI[30] and Arachne[31] and various conference presentations and publications by both Perseus and German team members. For example, project staff presented on the catalog and other linked data efforts at the two LAWDI (Linked Ancient World Data Institute) workshops.[32] In addition, two papers in particular by Perseus helped to broadcast the results of the Hellespont project.[33] Firstly, "Student Researchers, Citizen Scholars and The Trillion Word Library" published in the 2012 *Joint Conference For Digital Libraries* proceedings[34] explored the importance of the open Greek and Latin data collections that Hellespont funding helped to produce in terms of supporting student annotations and e-learning. Secondly, "Cataloging for a Billion Word Library of Greek and Latin," published in the 2014 proceedings of the *DATeCH '14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*[35] provides an overview of the linked data and technologies behind the Perseus Catalog and PDL's current CTS-URN implementation.

b. Accomplishments and "Failures" of the Hellespont Project

Perhaps the most significant accomplishment of the Hellespont project has been in investigating, approaching and documenting what the original narrative identified as "the problem of integrating the material and linguistic record of the past." To begin with, to even first investigate the problem of integrating these disparate records (the linguistic record of literary texts in Perseus and secondary works in Arachne and the material record of archaeological object data in both collections) first required the conversion of Greek and Latin texts in Perseus to the more up-to-date standard format of TEI P5 and the mapping of legacy object data into the standard format of CIDOC-CRM to support integration.

The original work plan and methodology/standards listed were largely followed with varying degrees of success. The fields between Perseus and Arachne were not aligned with each other but instead with the CIDOC-CRM model, as Arachne had already successfully completed a mapping of its database fields to this model. This process turned out to be far more problematic than even first expected, as outlined in further detail above. Additionally, initial exploration into the integration of TEI documents with CIDOC-CRM data as well as with automatically mined bibliographic information from sources such as JSTOR was conducted largely by the German

---

[30] http://hellespont.dainst.org/startpage/
[31] http://arachne.uni-koeln.de/drupal/?q=en/node/231
[32] http://wiki.digitalclassicist.org/Linked_Ancient_World_Data_Institute
[33] Full citations and links to open access versions of these publications is provided in section G. Grant Results.
[34] http://dl.acm.org/citation.cfm?doid=2595188.2595190
[35] http://dx.doi.org/10.1145/2595188.2595190

team through intense manual annotation and encoding of a sample text from Thucydides' *Historiae* known as the Pentecontaetia as a testbed virtual research environment.[36]

The conversion of all texts into TEI P5 Unicode is still a process that is ongoing, although one of the biggest lessons learned from this process was a goal articulated in the narrative: the grant enabled Perseus to extensively survey its large archive of Latin and Greek primary TEI texts and examine how these works were encoded over 20 years and to therefore "review, regularize, and document these practices and to provide textual sources that are easier for third parties to use, refine, and expand." Much of the work conducted for the Hellespont project has laid the groundwork for the creation of the Perseus canonical texts repository,[37] a publicly available Github repository that will ultimately serve as the home for all TEI texts and annotations published within the PDL.

Another area of accomplishment has been the successful development of standard and stable identifiers for much of the data within the PDL using CTS and CITE URNs.[38] Although plans to use automatic methods (such as named entity recognition) and community feedback to the system to automatically add markup and identifiers at the individual named entity and word level within texts have not been completed, work in this area is ongoing.[39]

Initial plans to represent "more sophisticated patterns about material culture" beyond TEI encoding of named entities such as temporal encoding through the use of TimeML were not able to be fully explored. The original plan had been to "tag an initial corpus of Greco-Roman sources to identify different patterns of thought across authors (Herodotus vs. Thucydides), time (Thucydides vs. Polybius vs. Procopius), genre (Herodotus and Thucydides vs. Pausanias), and language (Thucydides vs. Sallust, Livy, and Tacitus)" but this proved to be an overly ambitious goal for the project. A related deliverable was to support data mining and visualization of the markup methods to all of the new markup (either automatically added or hand created) and this goal was largely met through the adaptation and implementation of the GapVis tool for the encoded Thucydides text.[40]

---

[36] For more on this work, please see Romanello, Matteo, and Agnes Thomas. "The World of Thucydides: From Texts to Artefacts and Back." in Zhou, M., Romanowska, I., Zhongke, W., Pengfei, X., Verhagen, P. (eds.) *Revive the Past. Proceeding of the 39th Conference on Computer Applications and Quantitative Methods in Archaeology*, Beijing, April 12-16, pp. 276–284. Amsterdam University Press (2012) and Thomas, Agnes et al. (2013). "Insights in the World of Thucydides: The Hellespont Project as a Research Environment for Digital History." *Digital Classicist London & Institute of Classical Studies* seminar; August 9th, 2013. http://www.digitalclassicist.org/wip/wip2013.html; http://hellespont.dainst.org/startpage/files/wip2013-10at.pdf

[37] https://github.com/PerseusDL/canonical/

[38] For future plans of this work, see http://sites.tufts.edu/perseuscatalog/documentation/release-notes/next-steps-roadmap/roadmap-towards-linked-data-standards-compliance/

[39] Some work was conducted in this area in terms of the creation of treebanks to document various interpretations of words and entities such as the "syntactical and tectogrammatical trees of Thucydides' *Pentecontaetia* by Francesco Mambrini" is available for download at the Hellespont website.

[40] This work was succesfully completed by the German team at the DAI and Arachne and further information is available here: http://gapvis.hellespont.dainst.org/#index. Also see

One major goal of the original grant that was not fully accomplished was to create a "virtual collection" viewable online that fully integrated the textual and object data of the PDL with Arachne and also to make that data downloadable by third parties. While the original plan had been to make all data created available through three locations: 1) Perseus data through the Arachne interface, 2) Arachne data through the Perseus interface and 3) catalog data through DAI's Zenon bibliographic database, a growing focus on linked data within the digital humanities and library communities and movement away from tightly coupled integration of specific projects led us to consider different possibilities.

In some ways this "failure" also illustrates a shift in direction by Perseus in the last few years as our collections have scaled up greatly. More significant energy was devoted to making all of the data within Perseus fully open access and available as linked data[41] following standard citation practices for download by all interested parties rather than an exclusive focus on the integration of Perseus with another digital project such as Arachne, however significant. This is evidenced by the creation of a public GitHub repository for not just all code developed but for all Perseus texts (discussed above) as well as catalog data[42] and other datasets.[43] In addition, a searchable and browsable Linked Data interface[44] has been made available to provide access to the aggregated CIDOC-CRM data of the Perseus A&A data and the Arachne object database, although it has largely been implemented as a "proof of concept."

As was noted in the original narrative, "In the long run, the data that we produce for download will, we believe, be the most important product of our work, enabling third-party development and the emergence of new federated collections….," a deliverable that has on the whole largely been accomplished.

Another major desideratum of the original grant was to "address the problem of integrating the curated collections within Arachne and Perseus with the vast and rapidly expanding materials available in digital form from other sources" particularly in terms of providing training data for text mining services. At the time of the grant's writing the project was involved in separately funded grant projects, namely the NSF funded "Mining a Million Books Project"[45] directed by UMass Amherst and the eTraces[46] project centered at the University of Leipzig. The goal of

Romanello, Matteo. (2014). "Mining Citations, Linking Texts." *ISAW Papers* 7 (2014). http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/romanello/

[41] For more information regarding the production and plans for Perseus Linked data, please see Almas, B., et al. (2014). "Linked Data in the Perseus Digital Library." *ISAW Papers 7*: *Current Practice in Linked Open Data for the Ancient World,* New York : Institute for the Study of the Ancient World, New York University, http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/almas-babeu-krohn/

[42] As noted above, rather than making the catalog data available through Zenon, a separate searchable version of the catalog has been made available at http://catalog.perseus.org

[43] For a full list of repositories that the PDL maintains on GitHub, go to https://github.com/PerseusDL/.

[44] http://linkeddata.hellespont.dainst.org/). For more on the development on this interface, please see Schwane, K, et al. "The Hellespont Project: Integrating Arachne and Perseus in a new Linked Data Interface." Digital Classicist Berlin Seminar, December 17th, 2013. Prezi: http://prezi.com/emt9lxddpo5v/the-hellespont-project-integrating-arachne-and-perseus-in-a-new-linked-data-interface/

[45] http://ciir.cs.umass.edu/research/massivedata/

[46] http://etraces.e-humanities.net/

providing training data was partly accomplished through the release of various linked data sets described above that may be repurposed in different ways by other projects, and in particular results of this work have informed current work with the HathiTrust in terms of finding and making available more classical editions for the public domain as well as mining their collection automatically for more editions of classical texts.

In addition, another stated goal of the original grant was for Arachne and Perseus to "produce not only Greek/German and Latin/German parallel corpora but also a German/English parallel corpus of documents on classical studies." Due to both time and technical constraints, this was an area where significant development was unable to be undertaken. Similarly, the original grant noted that funding would allow Perseus "for the first time, to focus upon the problem of integrating scholarship in English and a major European language" with one major goal being significant research in terms of multilingual named entity detection for Greek, Latin, German, English and a related goal of increasing "multi-lingual access to scholarship in Classical studies in German and English."

Nonetheless, all of the work conducted under the auspices of the Hellespont grant as well as ideas proposed in the narrative that were not completely brought to fruition have helped to inspire the work of both the Open Philology Project and the Open Greek and Latin Project, two projects now being managed by PI Greg Crane as part of his responsibilities as the Humboldt Chair of Digital Humanities at the University of Leipzig.[47] These projects seek to not only increase the amount of classical texts in Latin and Greek available as open access and linked data but also to digitize translations, secondary and reference works regarding these primary texts in multiple languages to promote multi-lingual scholarship and teaching regarding the classical world.

c. Audiences

The audience for this project to begin with was largely existing users of the PDL and the Arachne database of the DAI. The PDL has an extensive existing user base with an average 250,843 visitors per month between July 2010 and June 2014, and its audience is reasonably diverse including students, scholars and members of the general public. The Hellespont project in particular, however, as it was focused on the creation of a virtual environment that combined textual and archaeological data from both collections to support advanced scholarship of necessity focused largely on the most advanced users of both projects. These users include professors of classical studies, philology, linguistics and other disciplines, archaeologists, and often independent scholars as well. This audience is also international in its scope. The Hellespont project largely demonstrated to this audience the possibilities inherent and importance of mapping and linking data between disparate digital projects as well as the need to be make all of this data available as linked open data.

d. Evaluation

The Hellespont project did not undergo any formal evaluations during the course of the grant. From our own assessment, the Hellespont project was reasonably successful in terms of accomplishing the intellectual ambitions if not always the concrete goals of the original proposal.

---

[47] http://www.dh.uni-leipzig.de/wo/

Certain tasks such as the mapping of the Perseus A&A database to the CIDOC-CRM, the only partially completed conversion of the Perseus Greek and Latin texts from betacode P4 to P5 Unicode consumed far more project time and resources than was originally anticipated.

One strength of this project was the collaboration and the collegiality of the project teams. Using multiple communication channels and visits by the PIs to the German and Boston locations respectively helped to support this project which required a great deal of collaborative efforts to produce its final results. Another strength of this project was the number of research outputs including multiple research publications, an aggregated CIDOC-CRM archaeological data set available in RDF, and the release of all the data included in the Perseus Catalog as well as other Perseus datasets on GitHub.

e. Continuation of the Project

Much of the work conducted during this project has informed the work of two other projects in which the PDL is participating, the Perseids Project and the Open Philology Project (as discussed above). In addition, the Perseids Project,[48] "a collaborative online environment in which users can edit, translate, and produce commentaries on a variety of ancient source documents" is making use of the catalog, A&A and textual data created as part of Hellespont and is further exploring issues of user annotation, collaborative data creation, detailed semantic encoding, and ways to make that data sharable and permanent.

The partnership between the Perseus Project and the DAI was strengthened during the course of this work. Indeed, as project PI Crane currently spends half his time in Leipzig there have been expanded opportunities for future collaboration with the DAI in terms of the creation of more open data in Greek and Latin and the ongoing investigation of an open and enhanced publication series for multi-lingual classical scholarship.

At present, the Open Philology Project, started by Crane at the University of Leipzig, constitutes a major continuation for the textual side of the Hellespont Project, with the catalog and textual data developed for Hellespont providing the foundation for the Open Greek and Latin component of Open Philology. In addition, Crane and Foertsch are planning to submit complementary proposals to the DFG in fall 2014, with one proposal focusing on the materials record led by Foertsch and another focusing on open data for Greek and Latin led by Crane.

f. Long Term Impact

The long term impact of this project will perhaps best be measured by the reuse of data and texts released through Github under its auspices.

g. Grant Products

Websites

Hellespont Project website
http://hellespont.dainst.org/startpage/docu.html#

---

[48] http://sites.tufts.edu/perseids/

Perseus Catalog
http://catalog.perseus.org

Publications

Crane, Gregory, Bridget Almas, Alison Babeu, Lisa Cerrato, Matthew Harrington, David Bamman, and Harry Diakoff. "Student researchers, citizen scholars and the trillion word library." *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. New York, NY, USA: ACM, 2012, 213-222. Open access preprint available at: http://hdl.handle.net/10427/75559

Crane, Gregory, Bridget Almas, Alison Babeu, Lisa Cerrato, Anna Krohn, Frederik Baumgart, Monica Berti, Greta Franzini, and Simona Stoyanova. "Cataloging for a Billion Word Library of Greek and Latin." *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. New York, NY, USA: ACM, 2014, 83-88. Open access copy available at: http://dl.tufts.edu/catalog/tufts:sd.0000171